may come from sparse NMR data or other experimental techniques, low-resolution protein structures can be reproducibly and rapidly assembled for proteins containing up to 250 amino acids or more.

The SICHO model employed in this example is very similar to that used in Example 1, although there are some differences in the protein representation that slightly increase the geometric fidelity of the model.

Reduced representation of polypeptide chains

The model chain consists of a string of virtual bonds connecting the interaction centers that correspond to the center of mass of the side chains and the backbone alpha carbons. All heavy atoms have the same weight in this averaging. Thus, the center of glycine coincides with its C_α, the center of alanine is located in the middle of the C_α-C_β bond, the center of valine roughly coincides with the C_β atom, etc. These interaction centers (beads) were projected onto an underlying cubic lattice with a lattice spacing of 1.45 Å. This constant defines the spatial resolution of the model. Obviously, the virtual bonds resulting from such a projection are of various lengths, depending on the identity of the two corresponding residues, the main chain conformation and the rotameric state of the side chain (*see* Figure 10). A change in any of these variables may change the corresponding virtual bonds (the chain vectors **v**). In proteins, these distances have a quite broad distribution, ranging from 3.8 Å for a pair of glycines to about 10 Å for some pairs of large side chains in their anti-parallel orientation and expanded conformations. The corresponding set of lattice vectors covers this distribution with good fidelity. The shortest vectors were of the form of (±2, ±2, ±1) or (±3,0,0) vectors, including all possible permutations. The length of these vectors corresponded to a distance of 4.35 Å. The longest lattice vectors were of the (±5, ±2, ±1) type and their length corresponded to 7.94 Å. Thus, the wings of the distribution are cut off. This should not have any noticeable effect on the model's fidelity because the small distance cut-off error is well below the resolution of the model, and the long-distance cut-off error is not important due to

very rare occurrences of distances above 8 Å. As a result, the set of allowed lattice bonds consists of 646 vectors, and sequentially adjacent vectors could not be identical.

A cluster of excluded volume points was associated with each bead of the model chain. Each cluster consisted of 19 lattice points: the central one; six points at positions (±1,0,0), (0,±1,0) and (0,0,±1) with respect to the central one; and 12 points at positions (±1,±1,0), including all permutations. Thus, the closest approach positions of another cluster with respect to a given cluster were of the form (±2,±2,±1) and (±3,0,0), as measured between the cluster centers. It could be easily calculated that, here, there were 30 closest approach positions. The distance of the closest approaches nicely corresponded to the smallest values of the inter-residue distances in real proteins. Since the average "contact distances" (*see* the following sections) of the model residues were somewhat larger than the distance of the closest approach, there were many more than 30 spatial orientations of two residues being in contact. Consequently, such a representation of protein structure avoided various anisotropy effects typically seen in the lower resolution lattice protein models. Figure 11 shows a small fragment of the model chain confined to the underlying cubic lattice with a lattice spacing equal to 1.45 Å. The excluded volume points are denoted by the solid and open circles. The solid circles indicate the three lattice points along the direction orthogonal to the plane of the figure: one in the plane below and one in front of the plane. The open circles denote points in the plane. With the above geometric restrictions, all PDB structures[3] could be represented with an average root mean square deviation (RMSD) of about 0.8 Å. Again, the accuracy of the fit does not show any systematic dependence on protein length nor on the orientation of the crystallographic structure with respect to the lattice coordinate system. Some features of the model chain are illustrated in Figure 10.

Conformational updating

The simplicity of the model protein representation facilitated the very rapid sampling of conformational space. The Monte Carlo algorithm employs three types of conformational transitions. The first type is a single bead, two-chain vector move. A random displacement of a randomly selected bead is generated and approved provided that the vector lengths and the excluded volume are not violated. The range of a random displacement is from 1 to $5^{1/2}$ lattice units. When accepted by the Metropolis criterion[4] (*see* the next section), such a move is equivalent to a collective rearrangement of the main chain and/or the side chain internal coordinates in a real polypeptide chain. The force field of the model, especially its generic components, prevented the acceptance of nonsensical, non protein-like conformations.[17] The second type of motion involved the permutation of three chain vectors. This was a larger scale move that was relatively rarely accepted due to possible steric interactions. The last type of move involved a randomly selected fragment consisting of several chain units. This fragment moved as a rigid body due to appropriate small changes in the two flanking chain vectors. For instance, such a move could translate a helical segment by a small distance, thereby slightly changing the conformation of the corresponding turn or loop regions.

Interaction scheme

The model force field consisted of several types of potentials. The first were generic biases that penalize against non protein-like conformations. These potentials were sequence independent. Sequence specific contributions to the force field consisted of knowledge-based two-body and multi-body potentials extracted from a statistical analysis of known protein structures. Finally, there were two kinds of potentials that contained evolutionary information extracted from multiple sequence alignments. In all cases, all PDB structures whose sequences were similar to the query sequence have been removed from the structural database used in the derivation of the potential (greater than 25% sequence identity).